DOCUMENT RESUME

ED 454 866 IR 058 157

| | |
|---|---|
| AUTHOR | McCallum, Sally |
| TITLE | Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives. |
| PUB DATE | 2000-11-00 |
| NOTE | 18p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144. |
| AVAILABLE FROM | For full text: http://lcweb.loc.gov/catdir/bibcontrol/mccallum_paper.htm l. |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | Access to Information; Bibliographic Records; *Cataloging; Library Role; Metadata; Standards; World Wide Web |
| IDENTIFIERS | Dublin Core; *Electronic Resources; *MARC |

ABSTRACT

This paper looks at three avenues of exploration related to bibliographic records that the World Wide Web environment invites--sorting out the level of control for Web material, reevaluating aspects of descriptive content requirements for these materials, and experimenting with new format structures. Part I discusses extending MARC for Web resources, including the extensive development of online electronic resources in the early 1990s, the establishment of a MARC field for electronic location and access, and key aspects of the Web's bibliographic control environment. Part 2 presents the following alternatives for control of Web resources: (1) unbundling the components of MARC, including MARC structure, content, and markup; (2) the Dublin Core data element set, including contributions of and issues with Dublin Core; (3) XML (eXtensible Markup Language) structure; and (4) the Resource Description Framework (RDF). Part 3 explores several related topics, including Web objects and the level of control, the "ephemeral" Web, the "research" Web, reevaluation of descriptive content, and exchange record structure. The conclusion suggests an agenda for librarians. The appendix presents a chart of the names for basic resource description metadata for common HTML (HyperText Markup Language) headers, Dublin Core elements, MARC core elements, TEI (Text Encoding Initiative) header elements, and ISO (International Standards Organization) 12083 elements. (MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 454 866

# Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives

Sally McCallum
Chief, Network Development and MARC Standards Office
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
B. Wiggins
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



Final version December 2000

## Introduction (1)

The Library community has been using the MARC format as a bibliographic data exchange structure for 30 years. Its data supports simple and complex retrieval by end users of information, it is the foundation of cost-saving copy cataloging, it is the underpinning to the proliferation of interchangeable and modular bibliographic control systems that have enabled libraries to automate in an integrated manner, it is the anchor around which a rich array of tools that help libraries do their work have been built, and it has become a language that thousands of bibliographic control staff use to input and discuss control issues. This usefulness has been built over the years as systems, tools, training and globalization made the MARC standard into a keystone for automation and development. While the MARC format is simply a communications format it turned out to be the key standard and has been used in innovative ways inside and outside of systems to provide users with retrieval and services unheard of 30 years ago.

During those 30 years information resources have also gone through evolutions. In the early MARC days the challenge was achieving consistent bibliographic control of textual material, then as cataloging standards for non-textual resources were developed, the format was constantly enhanced to accommodate them -- maps, music, graphics, moving images, etc. By the late 1970s the computer file became an important library resource and various forms of expression -- text, graphic, cartographic, and sound -- began to appear in electronic and digital forms as exemplified by the digital CD that has almost completely replaced the "vinyl" phonodisc. But an explosion took place in the last 10-12 years with the development of a communications vehicle for electronic data, the Internet, and subsequent breakthroughs in systems and software that established the web environment.

IR058157



BEST COPY AVAILABLE

Today's challenge is to provide appropriate tools for finding and retrieving the burgeoning web resources, and today's conference is looking at past practice and tools and their adaptability and applicability to the new environment.

## Part 1: Extending MARC for Web Resources

By the late 1970s libraries needed to control and provide access to machine-readable data files, "MRDF", along with other non-electronic resources. As a result, through the help of an American Library Association (ALA) committee of specialists, data elements were added to the MARC format to describe these files, indicate their sources, and provide for their standard numbers. With the advent of personal computers and CDROMs in the early 1980s additional elements were added, especially to indicate physical attributes and requirements of the media. At that time the new media were considered still to be in an evolving state, so there was caution about adding too much specialized descriptive information to the bibliographic record. What would have lasting usefulness for retrieval and deployment of the media was not well understood.

Then in the early 1990s came extensive development of *online* electronic resources, including gopher technology followed by the web. There was an immediate need to sort out description issues for these resources, and especially to provide electronic links from the bibliographic record to the actual resource -- which might be local or remote. In 1993, even before the Uniform Resource Locator (URL) addressing schema was completely developed, a MARC field (Electronic Location and Access, 856) was established for information what identified the path to a resource. That field has been adjusted at least annually since then as the Internet and web environments changed and matured -- the URL became an Internet Engineering Task Force (IETF) recommendation (a type of Internet standard); access methods expanded and had to be accommodated; file format types were better understood; and work on the Uniform Resource Number (URN) was completed. More recently, with the increasing numbers of web documents and the further development of linking to support navigation, the inclusion of URLs/URNs for related materials has meant allowing them to occur as needed in various MARC fields. This is an area where changes will continue to occur.

Spearheaded by the MARC format maintenance process, major discussions took place in the mid 1990s about the "real" nature of electronic information. Initially electronic data was treated like a new form of expression, first called "data files", then "computer files". As more material appeared in electronic form, the bibliographic community revised its views, changing terminology again to "electronic resources" and recognizing that in most cases computer files are a media that carries information in a recognized form of expression, such as textual, cartographic, graphic, etc. The MARC format was adjusted to be able to encode this view, enabling users retrieval of works across physical media, for example, a specific text in print, microfiche, and electronic form.

Recent work has identified a new "issuance" pattern for electronic resources that frequently change, coining the term "integrating" to describe them and identifying their differences from traditional serially

issued and monographic material. MARC discussions for accommodating the integrating resource model in the format are being held in parallel with the bibliographic discussions so when the community decides on the final requirements the format can quickly respond.

The discussions in the library community have thus focused on description and retrieval of electronic resources *along with* other physical information media, rather than separately. Distinguishing characteristics of electronically presented material are identified so that established cataloging principles for these other media can take electronic documents into account.

It is clear that the MARC format can provide a vehicle for the description of web and networked resources, and has kept up-to-date with developments in the medium. This is very positive and reassuring given the large community investment in MARC-based control -- the vast body of important non-electronic resources to which MARC is the key to interchange and the cost saving bibliographic services and tools that have been built on the standard format. But there are two factors that point to possible new directions. The first is the enormous and growing number of electronic resources and the impossibility of applying all of the current cataloging practices to them, and the second is the potential to unite electronic resources with cataloging data in new ways to assist retrieval. These make it important that the bibliographic community experiment with three key aspects of its bibliographic control environment:

- differentiation and selection of resources for levels of control,
- reevaluation of descriptive content requirements for cataloging, and
- the exchange record format structure.

## Part 2: Alternatives for Control of Web Resources

### Unbundling "MARC"

MARC is a generic name used for a bundle of components that come together to create MARC cataloging (or "metadata") records. These components are: structure, content and markup. The MARC format employees a standard structure to form a container for the cataloging content, which is controlled by a number of content standards. The markup (or tagging or content designation) is designed to be data identification oriented, often indicating the semantic relationships of elements of the cataloging content.

*MARC structure.* The underlying concrete syntax used for the MARC record is a simple "introduction - index - data" structure specified in the ANSI/NISO X39.2 and ISO 2709 standards (2). The standards dictate the length and some of the content of the introduction (the MARC Leader) and a few rules and options for construction of the index (the MARC Directory) and data fields. The MARC format implementation of ISO 2709 specifies a few additional rules about exactly how the index entries are to be constructed for the MARC format (tag length, starting character position, etc.) and how data fields are configured (positional fixed, subfielded variable length, indicators, etc.).

When the format was developed the driving forces were: to efficiently accommodate variable length data,

to enable easy selection of subsets of data elements, and to provide sufficient semantics (parsing and markup) to support data element identification that would open up many possibilities for retrieval, internal system record configurations, and data manipulation. The MARC format also needed to be able to accommodate various bibliographic data models. In the library context this meant data constructed according to various national or earlier cataloging rules, in addition to new bibliographic models in the future.

The MARC record structure has been constant for 30 years. This (along with stability of the tagging and the content rules) has been a strong factor in the proliferation of systems and services related to bibliographic operations and the enormous interchange of records currently occurring among systems. This structure is, however, interchangeable as is illustrated by the fact that many (most?) systems do not hold records in the MARC format structure but treat it as a communications format, the intended use. Even the "MARC displays" and MARC-centric input templates so common in automated systems are not actully MARC structure but a layout of the markup components of the record.

*MARC record content and markup*. The goal of the MARC record content is broad -- to describe many facets of a resource in support of multiple purposes. The content is formed by this multiplicity of uses, the descriptive standards, and the perceived needs for consistent retrieval. The possible content has grown over time as the functionality that the exchange record was expected to support grew. The content has also been driven by changes and differences in content rules, causing new elements to be defined to identify new ways of expressing information. The overriding purpose of the record has been general resource discovery, although precise identification, selection, acquisition, item control, and preservation are among other basic functions the record supports. The uses and relationships behind the data in the bibliographic record have recently been analyzed in the study *Functional Requirements for Bibliographic Records* (3), sponsored by the International Federation of Library Associations and Institutions (IFLA). These functional requirements, which are already a major factor in metadata work, are no doubt being described and analyzed by other papers at this conference.

The MARC record data content, is largely driven by external standards that have been collaboratively developed and widely adopted by the bibliographic community over many years -- the International Standard Bibliographic Descriptions (ISBDs), Anglo-American Cataloging Rules, 2nd edition (AACR2), Library of Congress Subject Headings (LCSH) and other subject thesauri, Library of Congress Classification (LCC), Dewey Decimal Classification (DDC), various name authority files, various official standards (e.g., ISBN, ISSN), and requirements for cooperative projects. The markup in the MARC format that identifies the data content and its relationships is thus largely determined by these external standards in conjunction with judgement on the amount of parsing and identification needed for a machine to perform the functions that users require. MARC content and markup can be very thin or very fat, but it is always under the control of the external content rules that is tries to support.

*Structure and content and markup* have been differentiated above because in looking at new or different ways to support retrieval of networked resources these components need to be considered separately. Their suitability for web resources, and the impact of change, and pathways for change have different possibilities.

The information universe has never monolithically used MARC format-based exchange records, fundamental a part as they may have played. The vast and important journal literature has been generally (well) controlled by highly automated special subject domain abstracting and indexing services. Archival materials have traditionally been described in hierarchical lists, called finding aids, that are separately constructed for each archival collection. The finding aid was recently "automated" with the development of the Encoded Archival Description (EAD) DTD, for SGML or XML encoding of these aids. However when considering web and networked resources, the subjects of this conference, the terms that come up most frequently as the keys to networked and web resource discovery are Dublin Core, XML, and more recently, RDF. The first is a data element set and the latter concern syntax and semantics.

**Dublin Core Data Element Set**

The Dublin Core is a short name for a collection of 15 data elements that have been identified as useful for identification of networked resources. Work on the Dublin Core was initiated at a conference at OCLC in Dublin, Ohio, in 1995, with a broad group of participants from the computer and library communities. These data elements were refined and finalized at follow-on conferences and via electronic participation.

The original goal was for a simple set of elements that, if included in headers to web documents, would increase the efficacy of web resource discovery tools such as the "web crawlers", and also serve as a basis for fuller description of the resources, as might be needed if a description were to be added to a library catalog or other special metadata listing. As with any standard, propagation was difficult and the inability to have the set widely adopted for the original purpose meant that use was redirected. As a result interesting experiments have been conducted that take detailed cataloging from multiple repositories and extract the Dublin Core subset of elements from them. These Dublin Core subsets are then merged, providing top level resource discovery across repositories.

The 15 data elements were specified with a minimal stipulation of content rules, in keeping with the original intent for simplicity and flexibility of use. But with use came the inevitable push to add new data elements and qualifiers for existing ones, entity relationship information, content rules, and a markup for the 15 basic data elements. This is not surprising to the bibliographic community where there is constant pressure to extend a data element set, such as MARC, to serve additional media, functions, and new user groups, all with special requirements in addition to the core needs. Through multiple annual meetings and email discussion, sets of qualifiers and additional content rules have recently been established for the original Dublin Core. They are to be used when finer refinement of the 15 data elements are needed. The reality is that use of Dublin Core up until now has usually required the establishment of locally defined qualifiers. The agreed-upon extensions should fulfill some of those needs, but if users continue to have requirements for more detail, it is recognized that local elements will be established and used.

The use of Dublin Core data elements in the OCLC Cooperative Online Resource Catalog (CORC) project, which tries to maintain an ability to convert records between the Dublin Core element set and content rules (or lack of them) and the MARC content and rules, has been challenging. The differences in

the content necessitated the extensive use of qualifiers with the Dublin core information in order to support retrieval compatible with full MARC content data. Some of the more interesting aspects of the CORC project are the special tools that have been developed to assist in automatically deriving cataloging data from the electronic resources themselves, and automatic checking of subject, classification and name authorities. These tools are not really related to either MARC or Dublin Core, however, but to the content standards and requirements of the bibliographic community.

"Dublin Core" thus refers to several things. (1) A basic set of 15 data elements for resource description with minimal content rules. The data elements are obvious enough that an author of a web document could often supply them without training. They are also common enough that they are a subset of data elements used in a variety of files and data bases, not just MARC, and can therefore be used for constructing meta meta files for first stop retrieval. (2) Dublin Core is also an officially expanded set of elements. The expansion is for qualifiers that refine the 15 basic elements and others that allow naming of the content rule used for the data. This form still does not mandate specific content rules. (3) "Dublin Core" (in quote marks) is also used to reference an input interface developed by OCLC where OCLC users can catalog resources (electronic or non-electronic). The input is via a special labeled template, called the Dublin Core template. The system attempts to impose a specific set of qualifiers and content rules to make the data compatible with data commonly found in a MARC record -- content rules that relate to AACR2, LSCH, DDC, etc., and various code lists. This system has a parallel input using MARC tagging. This appears to make the data as much MARC Core as Dublin Core.

Contributions of Dublin Core:

- widely recognized basic data element set, obvious and general enough that authors (or machines) can possibly supply them.
- through CORC and other projects, research on tools to automatically create Dublin Core data elements from electronic documents

Issues with Dublin Core:

- for the library community, insufficient consistency of data content, partly due to lack of content rules
- where content standards are specified or recommended, sometimes different from those commonly used in the library community.
- so basic that most applications need to define additional elements or subelements.

**XML Structure**

While Dublin Core is a set of data elements, XML (eXtensible Markup Language) is a data structure, comparable to the ISO 2709 data structure used by MARC. XML is actually a sub-structure possible under the more general data structure standard SGML (Standard Generalized Markup Language) which has a header followed by a simple repetitive tag-data form. SGML is specified in the ISO standard 8879

(4) and XML for web documents is a World Wide Web Consortia (W3C) recommendation. SGML was developed with the markup of text as the target, but has also proven useful as a programming tool. SGML has been used extensively in the publishing industry for textual material where generally corporations develop their own tag set under the structure, making interoperability impossible without first understanding the meaning of the tags.

*SGML/XML tag sets*. In the SGML environment a tag set with application rules is a Data Type Definition (DTD) (comparable to the MARC concept of format). ISO and other groups have tried to establish tag sets for the SGML structure that could be broadly used (similar to the establishment of MARC 21 as a broadly usable tag set for the ISO 2709 structure). Commonly used tag sets would allow easy interchange of marked up data and interpretation of the data without special intervention. ISO 12083 is one standard tag set targeted for relatively straightforward modern publications. The Library of Congress actually uses that tag set for the SGML markup of the MARC 21 documentation. ISO 12083 is widely used by publishers, but with a great deal of publishers-specific augmentation of the tagging specified in the standard. Another well publicized SGML markup is that of the Text Encoding Initiative (TEI). The TEI DTD was developed to make possible very comprehensive markup of textual documents -- markup that could support textual analysis of the documents. TEI has also met with some success, being used in many specialized text projects such as the Making of America projects sponsored by the National Science Foundaton. The TEI also influenced the DTD used by the Library of Congress for its American Memory digital projects.

But very importantly, HTML is a DTD that uses the SGML structure. It is a standardized tag set and is familiar to all as the markup predominant in the web environment. HTML has been enormously useful for documents to be displayed by web browsers because the tagging is display-oriented, focusing on the presentation aspects of a document, thus supporting display without construction of special style sheets.

The wide variation in the SGML tag sets developed, the complications of developing a complex DTD for each application, the success of HTML and desire to enhance it without going to more complex SGML structures -- along with the desire for SOME variability -- led to the development of XML as a SGML subset with special rules. XML does not require a formal DTD, just a scaled down "schema", or else a promise to be well formed. A document markup with a tag set defined for use in an XML structure should, for web purposes, be accompanied by a style sheet which will define its display to a browser. The style sheet enables the tagging to move away from the HTML presentation tagging to element identification tagging. XML does not itself specify a tag set that can be applied but a data structure open for definition of tagging that identify a given set of data elements, from general (e.g., Dublin Core) to detailed (e.g., full MARC data content).

The ISO 12083 and the TEI DTDs have now been specified in XML versions. Two other recent developments are using XML for tagging metadata that describes documents. One is the ONIX, a joint European and American publisher format for communicating book industry product information in electronic form. Products may be electronic books or printed material. Besides descriptive information, it contains data for the book selling function. The descriptive data could be a future source of cataloging data. The second development is the Open eBook initiative sponsored by the National Institute of

Standards and Technology (NIST). The Open eBook format specifies tagging for electronic book content, using HTML and XML tagging. The document description information specified for inclusion in the document are the simple Dublin Core data elements.

(Expected) positive contributions for XML

- structure very similar to that currently used (HTML) on the web -- XML has been endorsed for future use on the web.
- likely to be the structure for markup of many networked resources
- easy to establish a tag set, especially if DTD and schema concepts are not necessary for an application

Issues with XML

- if tag sets defined for use with XML structures are all different, interoperability is affected
- standards for schema and style sheets are still under development
- web use and widespread deployment still experimental and supporting tools still being developed (but at a rapid rate)

**RDF**

A new development, that is in its infancy still, is the Resource Description Framework (RDF). RDF is being developed by the W3C with the goal of making it a basic building block of the "semantic web", a manifestation of the web environment where the data is sufficiently related and marked up to support dynamically defining and exploiting new relationships. RDF holds a great deal of promise, perhaps some of it unattainable, but is certainly a path worth research. It provides a structured way to analyze relationships. RDF is not a concrete structure, but would logically use XML for document markup (it is itself being defined using an XML syntax) and would probably be open to externally defined content rules. It is, however, not ready for practical use but is currently an important research and development effort that may add understanding to resource description and become an important component in the future development of the Internet.

## Part 3: Explorations

**Web Objects and the Level of Control**

Studies are just beginning to be produced that analyze the types of resources found on the web, but speaking in general terms, much of the open access web contains material that would not be collected in a library for research purposes. A generous estimate might be that 5% of the resources available on the web are of permanent research value and should especially be saved, cataloged, and preserved. (This is referred to below as the "research web material") The large part remaining is largely business information, often with a marketing orientation. (This is referred to below, for convenience, as the "ephemeral web material".)

BEST COPY AVAILABLE

## The "Ephemeral" Web

Considering the ephemeral first, there are two basic concerns for this conference, current and future search and discovery. A presumption is made here that the current web is accessible and that past snapshots, or something comparable, of the web content are taken and stored for access in archives. For this body of material, simple resource descriptions are needed, and these descriptions are only feasible, given the vast number of documents in this class, if the resource creator takes some responsibility. This was the need recognized at the outset of the Dublin Core data element development effort -- to have a universally recognized simple set of data elements that authors capture in headers to their documents.

A simple "ideal" set of header elements with metadata about the document it sits in is needed. The elements need to be standardized as much as possible without layering too many form and content rules. Assurance is needed that the data elements and their tagging are carried over to newer markups (XHTML, XML, etc.) used for web documents.

A major question with expecting the author to include metadata is: Will the author take the time to supply it? This cannot be assured but fuller headers in only 50% of current web material would be a substantial improvement. There are numerous approaches to encourage authors to add the data. For example, an editor tool that the author can use to have the header automatically generated -- as best it can -- from the document content, and which the author can then correct. Encouragement by the library profession to major web sites to include a standard set of metadata as a requirement. Web indexers (crawlers) joining with librarians to promote awareness of the need for metadata and the benefits to both authors and users -- generally keeping the need and benefit alive and before those who can influence author behavior.

A major objection often voiced about author supplied descriptions for web documents is the tendency of some resource creators to engage in deceptive packaging -- supplying descriptive terms that will be popularly sought but do not apply to the resource. This can never be fully controlled, but a variety of efforts can mitigate it. Tools can compare content to author-supplied descriptors when web indexers skim from the metadata.

The Dublin Core set of elements are an obvious starting point for the endorsable set of basic elements. Appendix A compares the very basic elements commonly used today in HTML documents (Column 1) -- the metadata "hoped for" by popular web indexers -- with the rich Dublin Core set (Column 2). But it is also important to engage and obtain the concurrence of a wide spectrum of librarians, especially reference librarians. The many Dublin Core implementation experiments could provide data on how well the set works in retrieval. Also the MARC *content* needs to be a consideration when determining this set of descriptors. While use of the library community's content rules such as AACR2 would not be feasible for authors, content compatibility should be a maximized as far as possible as it will facilitate the variety of configurations in which this author-supplied cataloging may be useful. These will range from databases with only metadata harvested from electronic documents to catalogs that incorporate metadata related to selected web resources with non-web resources. The chart in Appendix A also gives a comparison (columns 3-5), using the simple Dublin Core as the basic match set, of the simple metadata that is



BEST COPY AVAILABLE

currently specified in MARC and two other widely used standard DTDs (TEI header and ISO 12083). MARC, TEI, and 12083 all contain markup for considerably more metadata than Dublin Core, but have reasonable overlap with the Dublin set.

**The "Research" Web**

The smaller proportion of web documents that are of primary importance for current and future research will generally benefit from richer metadata that supports more precise searching, since integration of those records into the catalogs of libraries is important. Libraries will be taking steps to assure access to these resources and provide for their preservation, and they will want to continue to offer catalogs that assist the user in finding all resources, irrespective of media. Here the author-supplied metadata would be useful as a starting point for cataloging following established content rules and containing more detail.

**Reevaluation of Descriptive Content**

If libraries continue their experiments to save snapshots of the web (providing retrieval through document-carried metadata) while focusing formal cataloging and preservation on the part of the web judged to be of lasting research value, are there changes still to be considered for the cataloging descriptions for these resources? There are complexities in the current content of the bibliographic record for which the time may be appropriate to consider whether they are necessary in today's environment. Experts at this conference are no doubt analyzing and providing recommendations concerning many important content issues related to the cataloging of web resources, so the following relates to a special content issue that affects any cataloging *format or DTD*: the large number of data elements that are considered necessary by librarians, thus are currently supported by MARC tagging.

*Intentional duplication*. The bibliographic record carries many data elements in duplicate. This is largely driven by the tradition of providing information both in transcription form (as it appears on the piece) and in a normalized form. There are many examples of this in the format, for example the transcribed author name as it appears on the piece and the inverted and normalized author name, and the transcribed place of publication and the coded place of publication. Another building block of the cataloging tradition is communicating descriptive information through natural language notes for easy display to and understanding by the user. Such information has been used for retrieval but to assure consistent retrieval the information is often also in the record in a controlled or coded form. Examples are the language note and language code, and notes that identify names associated with a work and the corresponding fields with normalized forms of those names.

This duplication is defended on bibliographic grounds. Transcription is an aid to the end user to precisely identify whether the item is the one sought and to librarians and their machines to help identify automatically duplicate resources and duplicate records. Notes are end user friendly and clarify the characteristics of an item in human-readable form, while the normalized and coded data assists retrieval, especially retrieval from large files. Coded sata generally transcends language differences and can be very important for "weeding" a retrieval set through search qualification.

*Multiplicity and granularity of data elements*. In addition to core fields, each form of expression (text, cartographic, music, etc.) has special characteristics that can be differentiated and identified. Over time a very large number of data elements have been defined for bibliographic records for recording these characteristics. Often structured elements with each subpart individually identified, instead of unstructured notes, have been adopted because of the possibilities for retrieval precision. When content rules for descriptive data specify data elements that are made of identifiable parts, these elements are often parsed and each part is identified, even though the need for retrieval may be questionable.

The MARC tagging has expanded to support identification of duplicate and granular data elements. Although through special tagging conventions, the MARC format creates dual purpose fields and avoids some duplication, most duplication is easier dealt with if simply tagged as specified. When adding tagging for elements to MARC, tests are carried out to evaluate the need for separately identified data elements (needed for indexing/retrieval? for special display?), but the many special interests served by library cataloging data often successfully justify individual identification.

What needs to be considered -- in the context of other recommendations presented at this conference -- is whether the characteristics of the electronic material are different so that some duplication is unnecessary? Do "title pages" or their analogs in electronic documents have enough stability to make transcription as useful as it is for print or object oriented publications? Are special normalized forms of some data still as critical or is research producing information identification and searching tools that require less rigor since the whole document content may theoretically be searched? Are display, retrieval, and sorting requirements different for web resources, indicating less need for specificity?

These descriptive issues are perhaps the most difficult to address, given the large number of purposes bibliographic records are constructed to support. Even if the bulk of web resources are controlled with a simple set of data elements, with the expectation of less precise but adequate retrieval, the numbers of resources receiving detailed cataloging is large. Rather than fitting the electronic resources into the existing mold, this is an opportunity to check and confirm or change some of our approaches to description for this and perhaps other types of material.

**Exchange Record Structure**

The third aspect of the current environment that needs to be addressed is the exchange record structure. As indicated, MARC record content can be separated from the MARC record structure, allowing the use of different structures for exchanging the same data. This is not often considered since the products and services that are based on the MARC exchange record have developed *because of* the relative stability and predictability of the actual exchange format structure (in addition to the content). With every decade (or less) preferred data structures, possible data structures, and fashionable data structures for electronic data have changed with the development of different internal computer system architectures, so it is a tribute to the profession that automation *and* exchange have been nurtured by separating communications from internal data structures and stabilize the former. The applications can take advantage of current trends without interrupting record exchange. One good reason why the community might want to

consider an alternative structure now is the apparent convergence of markup standards for the electronic document/web/Internet environment that may stabilize with XML. Cataloging data embedded in document headers or cataloging data exchanged for display through simple web browsers could be more efficiently used if transmitted in an XML structure with standard tagging and content.

In 1995 the Library of Congress, recognizing that the advent of full document markup had interesting potential for coordination with cataloging markup, gathered a small group of experts with experience with MARC, SGML, MARC in SGML, and electronic text markup. That group looked as a variety of aspects of the MARC format and made recommendations on how to treat them under the syntax rules of SGML. Out of that collaboration, the Library of Congress, with support from a contractor with special SGML expertise, produced SGML DTDs for the MARC record content and conversion programs that converts between the SGML/ISO 8879 and the MARC/ISO 2709 structures.

Since 1996, the Library of Congress has made available from the MARC 21 web site: an overview of the DTD requirements specified by the above group and two DTDs -- the Bibliographic DTD, incorporating the bibliographic, holdings, and community information format data; and the Authority DTD, incorporating the authorities and classification format data. Since early 1998, PERL script conversion utilities that convert both ways between the 2709 and 8879 structures have been freely available from the site, along with other tools for experimenting with the DTDs The DTDs have recently been specified also as XML DTDs and these DTDs will be available through the web site. The Library plans to keep these DTDs and tools up-to-date and in step with the markup standards, unless those standards become too volatile.

The experts group recommended that structural transformations be possible without loss of data. Thus, one characteristic of these MARC DTDs is that the XML tagging is MARC-like -- the tags are the same tags used within the MARC structure, with a little elaboration. For example, the tag for a title in MARC is "245" and in the XML MARC is "mrcb245". This tagging similarity is also the key to the simple structure conversion utilities. The following shows very brief MARC (Example 1) and XML (Example 2) record fragments for comparison.

Example 1 - MARC/2709

[245] (part of directory entry)
10$aData on the web:$bfrom relations to semistructures data and XML /$cSerge Abiteboul

Example 2 - MARC/XML

<mrcb245 i1="1" i2="0"/><mrcb245-a>Data on the web:</mrcb245-a>
<mrcb245-b>from relations to semistructures data and XML /</mrcb245-b>
<mrcb245-c>Serge Abiteboul </mrcb245-c>

One of the attractions of using XML is its possible use as an input and storage structure, in addition to a communications structure. While many librarians know the MARC tags as a shorthand for data element names, there may be applications where staff do not. For example, within the Library of Congress, MARC templates with full word tagging are used in special applications for creating basic MARC records. With XML, after moving into the structure it is not difficult to convert among tag sets, especially

if tag equivalencies are provided. Thus, the "mrcb245" could be converted to "m.title" if that were useful. That is another piece of an experimental tool set that the Library plans to make available.

Since XML has become popular, other XML versions of the format have been created as part of different projects indicating experimentation is taking place. With MARC-in-XML tools available, records for the "research" part of networked resources, for which full MARC content cataloging is warranted, can be either produced in XML, depending on the system, or easily converted from a MARC system to XML MARC for attachment to the XML document. This will provide a smooth pathway to what may be an eventual transition. If the XML data structure seems to have staying power -- and that is a real question given the nature and pace of change in web development -- with these tools the bibliographic community *will not have a revolution in its resource investment but an evolution.* This is important for an industry without surplus funds and with the need to keep its primary funding directed toward obtaining information resources themselves, not the conversion of catalogs.

**Conclusion**

This paper has discussed three avenues of exploration related to bibliographic records that the web environment invites -- sorting out the level of control for web material, reevaluating aspects of descriptive content requirements for these materials, and experimenting with new format structures. These explorations will take place with or without the participation -- or leadership -- of librarians, but they should not. Librarians need to have prominent roles in all explorations so that their cumulated knowledge and understanding of document control and discovery are built upon, not slowly rediscovered and reinvented.

As information specialists, librarians need to enhance their technical skills and collaboration skills, so they can work successfully with computer professionals, who will ultimately write the systems. As librarians they need to affirm the value of integrated access to research material -- electronic and non-electronic, different forms of expression, old material and new, etc. As responsible information servers they need to keep up with the directions technology is headed -- Will the web last? Will XML be superseded in a few years? Will there be constant costly change? What are retrieval innovations that influence record content? They need something like the following agenda.

* Apply their well honed resource selection skills to web resources, establishing general and feasible guidelines.
* For the mass of web resources, use simple descriptions and use "commercial" finding systems, but:
  - Advocate for simple document descriptions embedded in web resources.
  - Evaluate simple Dublin Core for that role, and submit for any well-justified changes.
  - Assist in development of helpful tools to improve such simple descriptions.
* For identified research material, use MARC content -- heavy to light as needed -- but:
  - Evaluate for possible unnecessary data elements complexities.
  - Experiment with structural transformations such as XML for the MARC content.
  - Assure that tools are readily available for conversion among structures.
* Keep the library community's understanding of the value of and commitment to standards by continuing

to work together on any changes to conventions and standards.

**Footnotes**

1. There are a large number of terms being used in the broader information community that often mean approximately the same thing, but relate concepts to the different backgrounds of the players. For example librarians are sometimes confused that metadata is something new and a replacement for either cataloging or MARC. Metadata is cataloging and not MARC. In this article terms based on library specialist terminology are used, with occasional use of alternative terms indicated below, depending on context. No difference in meaning is intended by the use of alternative terminology.. The descriptions of the terms are indicative, not strict.
   *cataloging data or cataloging content = metadata*
   - used broadly, in this context, for all data (descriptive, administrative, and structural) that relates to the resources being described.

   *content rules*
   - rules for formulation of the data including controlled lists and codes.

   *data elements*
   - the individual identifiable pieces of cataloging data (e.g., name, title, subtitle) and including elements that are often called attributes or qualifiers (since generally this paper does not need to isolate data elements in to subtypes).

   *relationships*
   - the semantics that relate data elements, e.g., name is author of title, title has subtitle.

   *content rules*
   - the rules for formulating data element content

   *structure = syntax*
   - the physical arrangement of parts of an entity

   *record*
   - the bundle of information that describes a resource

   *format = DTD*
   - a defined specification of structure and markup

   *markup = tag set = content designation*
   - a system of symbols used to identify in some way the following data.
2. ANSI/NISO Z39.2, *Record Interchange Format*, and ISO 2709, *Format for Data Interchange*. The two standards are essentially identical in specification. ANSI/NISO has a few provisions where the ISO standard is not specific, but there is no conflict between the two standards.
3. *Functional Requirements for Bibliographic Records*. IFLA Study Group on the Functional Requirements for the Bibliographic Record. Munich, Saur, 1998.
4. ISO 8879, *Standardized General Markup Language (SGML)*.

**Appendix A - Basic Resource Description Metadata**

| Common HTML Header metadata | Dublin Core element | MARC core element | TEI header element | ISO 12083 element |
|---|---|---|---|---|
| | **Identifier** | Electronic Resource Identifier (856 $u) | | |
| | **Format** | Electronic Resource Identifier (856 $q) | <extent> | |
| <title> | **Title** | Title (245 00a) | <title> | <title> |
| <meta name = "author"> | **Creator** | Added Entry (720 $a) | <author> | <author> |
| | **Contributor** | Added entry (720 $a) | <name> | <author> |
| | **Publisher** | Publisher (260 $b) | <publisher> | <pub> |
| | **Date** | Date of publication (260 $c) | | <date> |
| <meta name = "keywords"> | **Subject** | Uncontrolled subject (653 $a) | <keywords> | <keyword> |
| <meta name = "description"> | **Description** | Summary, etc. note (520 $a) | | <abstract> |
| | **Language** | Language note (546 $a) | <language> | |
| | **Type** | Genre (655 7$a) | | |
| | **Coverage** | General note (500 $a) | | |
| | **Source** | Linking entry (786 0 $n) | | |
| <ahref> | **Relation** | Linking entry (787 0 $n) | | |

BEST COPY AVAILABLE

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

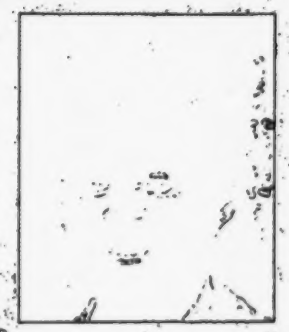
Logistical information for conference participants

Conference Organizing Team

# Bicentennial Conference on Bibliographic Control for the New Millennium

## *Confronting the Challenges of Networked Resources and the Web*

***sponsored by the Library of Congress Cataloging Directorate***



## Sally McCallum

Chief, Network Development and MARC Standards Office
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4160

## Extending MARC for Bibliographic Control in the Web Environment: Challenges and Alternatives

## About the presenter:

Sally McCallum is presently Chief of the Network Development and MARC Standards Office at the Library of Congress, the Office responsible for the maintenance of the MARC21 formats and a number of other interoperability-related standards such as an XML version of MARC, the Z39.50 Information Retrieval protocol, the Encoded Archival Description DTD, and the HTML standards used internally by LC for its web site. She has been an active participant in many organizations and working groups over her more than 20 years at LC, including the MARBI Committee of the American Library Association; boards and committees of the National Information Standards Organization (NISO); committees of the International Organization for Standardization (ISO) that develop standards for libraries and information services; and the Program for Cooperative Cataloging (PCC) She has also been very active in the International Federation of Library Associations and Institutions (IFLA), chairing the Professional Board and the Standing Committee on Information Technology and serving on format related committees responsible for the UNIMARC format. She has published a number of articles on standards and networking. McCallum has a BA from Rice University and an MLS from the University of Chicago.

Full text of paper is available

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

## Summary:

How will MARC accommodate changes to AACR2 and developments in alternative bibliographic control tools (DC, XML, RDF)? With the recent publication of *MARC 21*, MARC enters the new millennium as a proven and robust standard with a rich history of application in library OPACS and WebPACS worldwide. MARC was developed 30 years ago, long enough for the usefulness of a common format for data exchange to be appreciated and capitalized upon. Its broadly participatory maintenance process, well supported maintenance, and stability have enabled libraries to drastically cut cataloging costs AND to vastly enhance retrieval tools through automation of the catalog. But interoperability made possible by the format is ultimately dependent on the "interoperability" or compatibility of the data it carries. The cataloging conventions can make or break these savings and advances, and can be more critical than the actual carrier format.

This paper deconstructs the "MARC format" and similar newer tools like DC, XML, and RDF, separating structural issues from content-driven issues. Against that it examines the pressures from new types of digital resources, the responses to these pressures in format and content terms, and the transformations that may take place. The conflicting desires coming from users and librarians, the plethora of solutions to problems that constantly appear (some of which just might work), and the traditional access expectations are considered.